Molecular Evolution of the Plant *R* Regulatory Gene Family

Michael D. Purugganan*,[1] and Susan R. Wessler*,†

*Departments of *Botany and †Genetics, University of Georgia, Athens, Georgia 30602*

Manuscript received April 15, 1994
Accepted for publication July 18, 1994

ABSTRACT

Anthocyanin pigmentation patterns in different plant species are controlled in part by members of the *myc*-like *R* regulatory gene family. We have examined the molecular evolution of this gene family in seven plant species. Three regions of the *R* protein show sequence conservation between monocot and dicot *R* genes. These regions encode the basic helix-loop-helix domain, as well as conserved N-terminal and C-terminal domains; mean replacement rates for these conserved regions are 1.02×10^{-9} nonsynonymous nucleotide substitutions per site per year. More than one-half of the protein, however, is diverging rapidly, with nonsynonymous substitution rates of 4.08×10^{-9} substitutions per site per year. Detailed analysis of *R* homologs within the grasses (Poaceae) confirm that these variable regions are indeed evolving faster than the flanking conserved domains. Both nucleotide substitutions and small insertion/deletions contribute to the diversification of the variable regions within these regulatory genes. These results demonstrate that large tracts of sequence in these regulatory loci are evolving at a fairly rapid rate.

IT is widely believed that regulatory gene evolution is a significant factor in organismal diversification (DOEBLEY 1993; DICKINSON 1991). Changes in regulatory loci, for example, are thought to underlie the evolution of developmental mechanisms that result in morphological differentiation between taxa (GOULD 1977). Despite their central relevance to organismal evolution, however, relatively little is known about the molecular evolution of regulatory genes. Information on the patterns and rates of DNA sequence variation for regulatory genes is crucial if we are to understand the evolution of genes that control pattern formation in eukaryotes.

The genetic loci that specify the diverse anthocyanin pigmentation patterns in plants provide a system for studying the molecular evolution of regulatory genes. Anthocyanins are responsible for the purple-red pigmentation of the vegetative and floral organs of a large number of plant species. These plant pigments are believed assist in pollinator attraction, fruit dispersal and possibly UV protection (EPPERSON and CLEGG 1987; STAPLETON 1992). Genetic and molecular analyses in both *Zea mays* (maize) and *Antirrhinum majus* (snapdragon) have demonstrated that patterns of tissue-specific anthocyanin pigmentation are partially controlled by a gene family whose products regulate the structural genes of the anthocyanin biosynthetic pathway (LUDWIG *et al.* 1989; LUDWIG and WESSLER 1990; RADICELLA *et al.* 1991; GOODRICH *et al.* 1992). Members of this regulatory gene family in maize include the *R* and *B* loci, which direct pigmentation of inflorescence, leaf, root and seed tissues. Closely related but distinct *R* genes, including *R-S*, *Sn* and *Lc*, are found on chromosome *10*. These genes are believed to have arisen through the recent duplication of a single ancestral gene within the genus *Zea* (ROBBINS *et al.* 1989). A paralogous gene, *B*, is located on chromosome *2*, and appears to have arisen earlier during the evolution of the grass (Poaceae) family (ROBBINS *et al.* 1989; HELENTJARIS *et al.* 1988). Genetic mapping studies show that the *R* and *B* loci in *Z. mays* reside in chromosomal segments which apparently duplicated during a polyploidization event that produced the amphidiploid genomes of many grass species (AHN and TANKSLEY 1993; HELENTJARIS *et al.* 1988).

Genetic and molecular studies demonstrate that the maize *R* and *B* genes are functionally equivalent; any *R/B* gene is capable of activating expression of the structural genes in the anthocyanin biosynthetic pathway. Homologs to these maize genes are also believed to function as anthocyanin pigmentation regulators in other flowering plants. A homolog of the maize *R/B* genes, for example, has been isolated in the dicot plant species *A. majus*. This gene, *delila*, regulates the purple coloration of the snapdragon flower (GOODRICH *et al.* 1992). The maize *R* and *B* and the Antirrhinum *delila* genes constitute a regulatory gene family which we collectively refer to as the *R* family.

Molecular studies reveal that the genes in the *R* family encode *myc*-like proteins which contain a basic helix-loop-helix (bHLH) motif found in other eukaryotic transcriptional activators such as Max and MyoD1 (DEPINHO *et al.* 1987; DAVIS *et al.* 1987). Although all *myc*-like regulatory proteins share this conserved bHLH domain, they posses little sequence similarity outside this motif. The bHLH domain is responsible for the DNA binding activity of several transcriptional

[1] Present address: Department of Biology-0116, University of California at San Diego, La Jolla, California 92093.

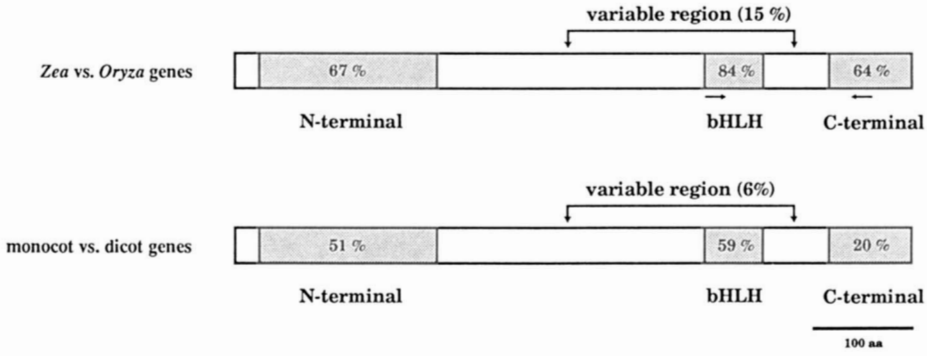

FIGURE 1.—Conserved and variable regions between *R* homologs. Shaded boxes delimit conserved domains; the percentage amino acid identities between the aligned Zea *Lc*, *R-S*, *B*, *A. majus delila* and Oryza *R* homologs (S. R. WESSLER, B. ANDERSON and J. P. HU manuscript in preparation) are shown. Variable regions flanking the bHLH domain are also indicated. Comparisons are depicted for genes in all three species, as well as those found only within Zea and Oryza. Horizontal arrows show approximate locations of PCR primers used to amplify *R* homologs from different *Poaceae* species

activators. Moreover, recent x-ray crystallographic studies reveal that the amino acid residues in the bHLH domain which interact directly with its cognate DNA target site are conserved between most *myc*-like proteins (FERRE-D'AMARE *et al.* 1993; ELLENBERGER *et al.* 1994).

In this report, we describe the molecular evolution of the *R* gene family in plants. This represents the first detailed molecular evolutionary analysis of a plant regulatory gene. We show that large regions of the *R* coding sequence are undergoing rapid evolution. This pattern of sequence divergence suggests that either most of the protein structure is under little functional constraint or that selection is acting to diversify the products of these regulatory loci.

MATERIALS AND METHODS

Sequence analysis of *R/B* family: Alignments of both nucleotide and protein sequences for the Zea *Lc*, *R-S*, *B*, *A. majus delila* and the Oryza (rice) *R* homolog genes were made using PILEUP of the UWGCG package. The nucleotide substitution rates were calculated by the method of LI *et al.* (1985), with a weighted computation method that gives unbiased estimates of synonymous and nonsynonymous substitution levels. Nucleotide substitution calculations were done using using the program LI93 (LI 1993). The maize (accession nos. M27227 and X57276) and Antirrhinum sequences (M84913) were obtained from GenBank. The rice sequence was provided by BETH ANDERSON (University of Georgia).

Molecular analysis of grass *R* genes: *Z. mays* and *Tripsacum australe* genomic DNA were isolated from leaf tissue using standard protocols (DELLAPORTA *et al.* 1983). Other grass genomic DNA samples were provided by L. ARTHUR (Pennisetum) and E. FRIAR (Phyllostachys). J. BENNETZEN provided Sorghum λ clones that were isolated from a genomic library using the maize *Lc* gene as a probe.

Grass *R* sequences were isolated from genomic or λ clone (Sorghum) DNA using the polymerase chain reaction (PCR). Degenerate PCR primers 1 (5′-GGGAACGGCAARAANCAYGTNATG-3′) and 2 (5′-AGGTGCRTCRAANACNCGKGTNATG-3′) were used; these primers are based on amino acid sequences in the bHLH and C-terminal domains which are conserved between Zea *R*, *B* and Oryza *R* genes. PCR reactions contained 200 ng of genomic DNA (5 ng for clones) and 700 ng of primers; buffer conditions were as described (VARAGONA *et al.* 1992). Reactions were conducted for 40 cycles at 95° for 1 min, 47° for 1 min and 72° for 2 min, followed by a 72° incubation for 10 min. PCR products were run on a 1.2% agarose gel and blotted on nylon filters (GeneScreen Plus). The presence of *R*-homologous sequences were confirmed by hybridization of PCR blots with a radioactively labeled *Lc* probe containing the bHLH and C-terminal domain sequences. PCR products were cloned using the TA cloning procedure (InVitrogen). Several clones were isolated from each species and sequenced with an automated sequencer at the University of Georgia Molecular Genetics Instrumentation Facility. Each clone was sequenced at least four times.

Sequence analysis was conducted using both the Intelligenetics and UWGCG packages. Alignment of nucleotide sequences was done by GENALIGN in the Intelligenetics package and refined visually taking both amino acid and nucleotide sequences into consideration; alignment used in Table 2 analysis included *delila* (not shown). Nucleotide sequences have been deposited in GenBank (accession nos. U11449 to U11451). The *R* homolog gene tree was inferred from nucleotide sequences by parsimony analysis using PAUP (SWOFFORD 1993). The tree construction was done with the PAUP heuristic search algorithm and with branch swapping performed using the tree-bisection-reconnection procedure. The PAUP MULPARS option was in effect. The Antirrhinum *delila* gene was used as an outgroup.

RESULTS AND DISCUSSION

Patterns of conservation and variability within the *R* coding region: Analysis of sequences between *R* family members from Zea, Oryza and Antirrhinum indicate considerable heterogeneity in levels of conservation between protein sequences. Figure 1 diagrams the regions of conservation and variability between the proteins encoded by the maize *R* and *B* genes, the *Oryza sativa* (rice) *R* homolog gene and Antirrhinum *delila*. The predicted protein sequence alignments reveal three domains that are conserved between these different monocot and dicot *R* homologs: (i) a centrally located 57-amino acid bHLH domain, with 59% amino acid identity between predicted proteins from all three species, (ii) a 175-amino acid conserved region at the N terminus of the protein which has 51% identity, and (iii)

TABLE 1

Levels and rates of replacement substitutions between different *R* homologs

	Nonsynonymous substitutions (Ka)			Replacement rate		
	Conserved	Variable	Total	Conserved	Variable	Total
Within Poaceae						
Zea *Lc*/Zea *B*	0.05 (0.02)	0.26 (0.03)	0.13 (0.01)	NA[a]	NA	NA
Zea *Lc*/Oryza *R*	0.16 (0.03)	0.68 (0.05)	0.35 (0.02)	1.2	5.2	2.7
Zea *B*/Oryza *R*	0.16 (0.03)	0.66 (0.06)	0.33 (0.02)	1.2	5.1	2.5
Monocot/Dicot						
Zea *Lc*/*A. majus delila*	0.36 (0.06)	1.24 (0.10)	0.65 (0.03)	0.9	3.1	1.6
Zea *B*/*A. majus delila*	0.36 (0.06)	1.47 (0.15)	0.64 (0.03)	0.9	3.7	1.6
Oryza *R*/*A. majus delila*	0.38 (0.06)	1.32 (0.12)	0.64 (0.03)	0.9	3.3	1.6

Number of nonsynonymous substitutions per site (Ka) and rates of replacement substitutions ($\times 10^9$ substitutions/site/year) are given for pairwise comparisons between Zea *Lc* and *B*, *A. majus delila* and Oryza *R* genes. Standard errors for Ka are given in parentheses. The calculations were done based on the alignment of the entire genes (see Figure 1).

[a] Not applicable.

a weakly conserved C-terminal domain with 20% identity. The conserved N-terminal and central bHLH regions correlate well with portions of the *R/B* proteins believed to be necessary for transcriptional activation. The bHLH domain contains the presumed DNA-binding and subunit dimerization activity of the *R* proteins, while the transcriptional activation domain may reside in the conserved acidic N-terminal region. Deletion analysis of the Zea *B* gene suggests, moreover, that the N-terminal region of *B* may also be involved in interactions with the *myb*-like *C1* gene of maize (GOFF*et al.* 1992).

Interestingly, the conserved C-terminal domain is much more weakly conserved (20%) than either the N-terminal (51%) or bHLH domains (59%) when comparing the Antirrhinum/Zea, Oryza *R* genes. The level of conservation in the C-terminal domain, however, is comparable to either the N-terminal or bHLH regions (64% *vs.* 67% and 84%, respectively) when only grass genes are compared. This suggests that the C terminus of the *R* genes are conserved between more closely related taxa but diverge significantly between evolutionarily distant species.

The conserved regions only account for approximately one-half of the structure of the *R/B* proteins. The rest of the protein is comprised of two large poorly conserved tracts of sequence. These two variable regions, 264 and 64 amino acids in length, flank the central bHLH domain. Together, these two variable regions share only 6% amino acid identity between *R* homologs from Zea, Oryza and Antirrhinum. Both replacement and insertion/deletion mutations contribute to the variation in these weakly conserved regions.

The increased variability within large regions of the *R* homolog genes are reflected in the levels and rates of DNA sequence evolution in the conserved and variable domains (Table 1). The replacement rate for these proteins may be calculated from the levels of nonsynonymous nucleotide substitutions (Ka) and the time since common ancestry between the maize, rice and Antirrhinum genes. Zea and Oryza last shared a common ancestor approximately 65 million years ago (mya) (CREPET and FELDMAN 1991), while the monocots and the eudicots diverged approximately 200 mya (WOLFE *et al.* 1989). Based on these divergence times, the mean replacement rate between *R* homologs is calculated as 2.0×10^{-9} substitutions/site/year. Replacement rates for the conserved domains (bHLH plus N and C terminus) average at 1.02×10^{-9} substitutions/site/year. The variable regions, however, have a mean replacement rate of 4.08×10^{-9} substitutions/site/year; this rate represents a nearly fourfold increase over the nonsynonymous substitution rate in the conserved regions. In contrast, the rate of synonymous nucleotide substitutions are roughly equivalent between the conserved and variable regions of the *R* genes. Based on comparison of the *R* rice homolog and the Zea *R* and *B* genes, the mean levels of synonymous substitutions (Ks) are 0.98 for the conserved domains and 1.43 for the variable regions. Based on the rice-maize divergence time, this gives a synonymous substitution rate of 7.5×10^{-9} substitutions/site/year and 11.0×10^{-9} substitutions/site/year for the *R* conserved and variable regions, respectively. It is apparent that the conserved and variable region synonymous substitution rates are nearly equivalent, and significantly higher than the *R* conserved domain nonsynonymous nucleotide substitution rates. Moreover, the levels of amino acid-changing nucleotide substitutions suggests that more than half the sequence of the *R* regulatory proteins are diverging at a fairly rapid rate.

The *R/B* genes in the Poaceae: To investigate the evolution of the conserved and variable regions of the *R* genes within the Poaceae, we isolated and characterized eight *R* homolog sequences from the grass species *Tripsacum australe* (gama grass), *Sorghum bicolor*, *Pennisetum glaucum* (pearl millet) and *Phyllostachys acuta* (woody bamboo). The sequences isolated contain the bHLH domain, the 3′-flanking variable region, and the upstream portion of the conserved C-terminal domain (see Figure 2).

```
                1                                                  58  59                76
Zea Lc          KNHVMSERKRREKLNEMFLVLKSLLPSIHRVNKASILAETIAYLKELQRRVQELESSR  EP-ASRPSETTTRLITRP
Trip R          KKHVMSERKRREKLNEMFLVLKSLLPSIHRVNKASILAETIAYLKELQRRVQELGSSR  EP-ASGPSETTTRLITRP
Zea B           KNHVMSERKRREKLNEMFLVLKSLVPSIHKVDKASILAETIAYLKELQRRVQELESRR  QG----------------
Sor R1          KNHVMSERKRREKLNEMFLILKLLVPSIQKVAKVSLLAETIAYLKELQRKVQELKSSR  EL-LSRPSETTARP-TKP
Sor R2          KKHVMSERKRREKINEMFLILKLLVPSIHKVDKASILTETIAYLKELQRGVQELESSR  EL--TTPSETTTRT-TRP
Pen R1          KKHVMSERKRREKLNEMFLVLKSLVPSIHKVDKASILAETIAYLNELQRRVQELESSR  EPMMLRQSETR--KVTR-
Pen R2          KKHVMSERKRREKLNEMFLVLKSLVPSIHKVDKASILAETIAYLKELQRRVQELESSR  EPMISRPSETR--KVTR-
Pen R3          KKHVMSERKRREKLNEMFLALKSLVPSIHKVDKASILAETIAYLKELQRRVQELESSR  EPMISRPSETR--KVTR-
Pen R4          KKHVMSERKRREKLNEMFLVLKSLAPSIHRMDKVSILAQTIAYLKDLQRRVQELEYSR  EPIISRPSETT--KVAR-
Phyl R          KKHVMSERRRREKLNEMFLILKSLVPSIHKVDKASILAETIAYLKELEQRVEELESNR  EP--SRPSETRGRR----
Oryza R         KNHVMSERRRREKLNEMFLILKSVVPSIHKVDKASILAETIAYLKELEKRVEELESSS  QP-SPCPLETRSRR----
```

bHLH Domain

```
                77                                                120 121               152
Zea Lc          SRGNNESV-RKEVCAGSKRKSPELGRD---DVERPPVLTMDAGT  SNVTVTVSD-KDVLLEVQCRWEELLMTRVFDA
Trip R          SRGNNESV-RKEVCAGSKRKSPELGRD---DVERPPVLTMDAGT  SNVTVTVSD-KDVLLEVQCRWEELLMTRVFDA
Zea B           -GSGCVSK-KVCVGSNSKRKSPEFAGG---AKEHPWVLPMD-GT  SNVTVTVSD-TNVLLEVQCRWEKLLMTRVFDA
Sor R1          CGIGSESV-RKKLSAGSKRKSPDFSGD--VEKEHPWVLPKD-GT  SNVTVAVSD-RDVLLEVQCRWEELLMTRVFDA
Sor R2          RGISNESA-RKKLCAGSKRRPSPEVGGDVVNKEHPWVLPKD-GT  SNVTVTVAN-TDVLLEVQCRWEELLMTRVFDA
Pen R1          RHDDDEDV---GN?SGSKRKASELGSG--VEREHP---TKD-DT  SNVTVTISN KEVLLEVQCRWEELMMTSVFDA
Pen R2          RHDDDEDV---GNGSGSKRKASELGSG--VEREHP---TKD-DT  ?NVTVTISN-KEVLLEVQCRWEELMMTRVFDA
Pen R3          RHDDDEPV---TKGSGSKRKGSELGSG--VAREHP---TKD-DT  TNVTVTISN-KEVLVEVQCRWKELMMTRVFDA
Pen R4          RHDDDEAVTRKVCAAGTKRKDSELSSD--VEREHPWEISKD-GA  SNVTVTVAD-KEVLVDVQCRWEELMMTRVFDA
Phyl R          RHEIAGIS-------GAKRKASSEPGGRDVERERLWALSMD-GP  SNVNVTVMD-KEVLLEVQCGWKEMLMTRVFDA
Oryza R         KCREITGK---KVSAGAKRKAPAPEVASDDDTDGE---RRH-CV  SNVNVTIMDNKEVLLELQCQWKELLMTRFVDA
```

C-terminal Domain

FIGURE 2.—Alignment of predicted protein sequences from *R* homolog sequences isolated from various grasses. The conserved bHLH and the 5′ end of the C-terminal domains are indicated. The variable region is located between these two conserved regions. The different source species for the various *R* homologs are as follows: Zea *Lc* and Zea *B* (*Z. mays*), *Trip R* (*T. australe*), *Sor R1, R2* (*S. bicolor*), *Pen R1-R4* (*P. glaucum*), *Phyl R* (*P. acuta*), Oryza *R* (*O. sativa*). The position of an intron found within the bHLH coding region of all grass *R* homologs is indicated by an arrow. Dashes indicate gaps. The indicated conserved bHLH/C terminal domains, and the intervening variable region were used to calculate the conserved and variable region Ka/Ks values in Table 2, respectively.

Only one *R* homolog sequence isolated from Tripsacum and Phyllostachys, while two and four independent *R* genes were isolated from Sorghum and Pennisetum, respectively. The sequence of these *R* homologs, as well as sequences from previously isolated Zea and Antirrhinum genes, were used to construct a gene phylogeny of the *R* family in the Poaceae (Figure 3). In general, this tree agrees with the accepted phylogeny of the grasses (DOEBLEY *et al.* 1990; E. KELLOGG, personal communication) confirming that these sequences are homologous to each other and to the maize *R* and *B* genes. The phylogeny indicates that the maize *R* and *B* loci shared a common gene ancestor around the time the Zea and Sorghum lineages diverged. This suggests that the polyploidization event that led to the amphidiploid Zea genome also occurred during this time. Interestingly, this tree also reveals that the Oryza and Phyllostachys *R* genes do not form a monophyletic group and that the Oryza *R* gene is basal to the rest of the grass *R* homologs. If the *R* gene tree accurately reflects the phylogeny of the Poaceae, then it may be that the Oryzoids are the true basal members of the grasses.

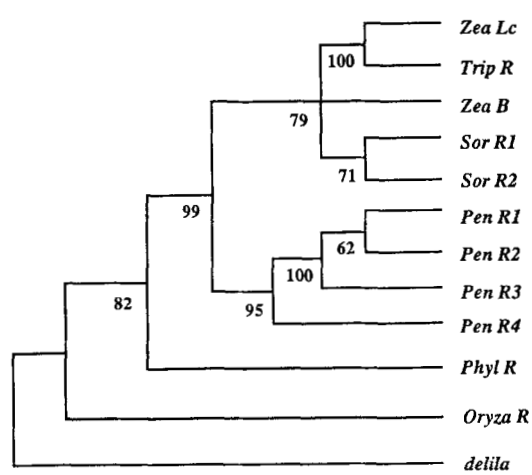

FIGURE 3.—Phylogenetic tree of *R* homolog sequences isolated from different *Poaceae* species. The tree was constructed under maximum parsimony from nucleotide sequences using PAUP. A total of 474 nucleotide sites within the coding sequences were used in the analysis. The numbers next to the nodes give bootstrap values from 100 replicates. The tree length is 809 steps, with a consistency index of 0.73.

TABLE 2

Pairwise comparisons of nonsynonymous/synonymous substitutions (Ka/Ks) ratios for *R* homologs within the Poaceae

Genes	Zea *Lc*	Trip *R*	Zea *B*	Sor *R1*	Sor *R2*	Pen *R1*	Pen *R2*	Pen *R3*	Pen *R4*	Phyl *R*	Oryza *R*
Zea *Lc*	—	NA[a]	0.94	0.79	0.65	0.84	0.72	0.71	0.83	0.66	0.64
Trip *R*	0.11	—	0.94	0.82	0.65	0.87	0.74	0.74	0.86	0.69	0.64
Zea *B*	0.09	0.09	—	0.48	0.48	0.74	0.74	1.08	0.47	0.90	2.64
Sor *R1*	0.18	0.29	0.17	—	0.77	0.75	0.70	0.71	0.59	0.57	0.77
Sor *R2*	0.13	0.15	0.12	0.49	—	0.69	0.66	0.57	0.67	0.74	0.83
Pen *R1*	0.08	0.10	0.06	0.17	0.07	—	NA	4.85	0.69	0.53	0.88
Pen *R2*	0.05	0.05	0.04	0.13	0.06	0.06	—	3.45	0.60	0.50	0.86
Pen *R3*	0.06	0.06	0.05	0.13	0.08	0.09	0.07	—	0.90	0.65	0.94
Pen *R4*	0.10	0.11	0.16	0.20	0.18	0.14	0.14	0.13	—	0.74	0.83
Phyl *R*	0.10	0.14	0.10	0.14	0.10	0.14	0.10	0.07	0.19	—	0.68
Oryza *R*	0.11	0.11	0.10	0.12	0.09	0.10	0.07	0.06	0.13	0.11	—

Ka/Ks ratios for the variable and conserved regions of the Poaceae *R* homolog sequences are given above and below the diagonal, respectively. These calculations were based on the sequence alignments depicted in Figure 2.

[a] Not applicable.

Evolution of the *R* variable domain in the grasses: Analysis of the isolated grass *R* homolog sequences reveal that several mostly conservative amino acid replacements between genes occur within the conserved sequences containing the bHLH and C-terminal domains. In the bHLH domain only 18 amino acid replacements occur between the 11 grass *R* homologs. Of the replacements found in the core bHLH region, only one (N/D_{32} to A_{32} in *Sor R1*) can be considered a highly radical replacement. Moreover, those amino acid residues that have been shown in another bHLH-containing protein to participate directly in DNA or dimerization contacts are invariant between all Poaceae *R* homologs (ELLENBERGER *et al.* 1994).

The variable region between these two domains reveal a significantly greater number of protein sequence alterations between genes. Table 2 shows the ratio of nonsynonymous and synonymous nucleotide substitutions (Ka/Ks) for the conserved (bHLH plus C-terminal) and variable regions between isolated grass *R* homolog sequences. The Ka/Ks ratio between these *R* homologs range from 0.04 to 0.49 within the conserved domains, with a mean value of 0.12. The Ka/Ks ratios for the variable region are consistently higher for all the *R* homologs within the Poaceae, ranging from 0.47 to 4.85. The difference in Ka/Ks values for conserved *vs.* variable domains are significant ($P < 0.001$) under a paired-sample *t* test. The high Ka/Ks ratios within these variable regions are evident even between *R* genes whose alignments are relatively unambiguous. Moreover, genes that group together phylogenetically also have high variable region Ka/Ks values, suggesting that the elevated levels of amino acid replacements observed are not merely due to large evolutionary separations.

The mean Ka/Ks value for the variable region is 0.89, a nearly 8-fold increase when compared to the Ka/Ks for the conserved region. In comparison, the mean Ka/Ks ratio for eight monocot and dicot genes is 0.14 (HUANG *et al.* 1992; MARTIN *et al.* 1989). Among 42 mammalian sequences, the mean Ka/Ks value is 0.189 (LI *et al.* 1985). The conserved regions of the *R* homologs analyzed appear to possess Ka/Ks ratios closer to the value found in most eukaryotic coding sequences. The variable regions within the *R* genes, however, fix replacement mutations at a much higher rate.

In general, the Ka/Ks ratio measures the contrasting tendencies of a protein toward functional conservation and evolutionary divergence. Most replacement mutations are subject to purifying selection, resulting in relatively low Ka/Ks ratios for most coding sequences (<0.2) (LI *et al.* 1985). An increase in Ka/Ks values may reflect the lack of sequence constraint on the protein. The Ka/Ks value for the variable region is close to one, suggesting that this region is evolving neutrally and at the rate expected for a pseudogene. Elevated nonsynonymous substitution rates in several genes, however, have also been attributed to positive selection (HILL and HASTIE 1987; HUGHES and NEI 1988; TANAKA and NEI 1989; HUGHES 1991). It remains to be determined whether selection is operating to diversify the structure of *R*. It is possible that the rapid divergence of variable regions may contribute to species- and even gene-specific differences in regulatory activity (WHITFIELD *et al.* 1993; TUCKER and LUNDRIGAN 1993). Transient assays have demonstrated, for example, that the maize *B* gene and rice *R* homolog are less effective than maize *Lc* in activating the maize *bronze* promoter in aleurone tissue (R. DAMIANI and S. R. WESSLER, unpublished observations). The greater activity of *Lc* in maize tissues compared to either the *B* or rice *R* homolog may arise from differences between the variable regions of these genes.

Another regulatory gene was recently shown to possess the same pattern of molecular evolution as the *R* family. The mammalian *SRY* sex-determining genes were reported to contain regions of rapid sequence evolution flanking a conserved DNA-binding domain (WHITFIELD *et al.* 1993; TUCKER and LUNDRIGAN 1993). Interspecific variation between Murinae *SRY* arises in part from differences in the length and composition of simple repeat motifs found at the C termini of these

genes. Repeat motifs are also partly responsible for rapid variation in portions of the Plasmodium circumsporozite protein (HUGHES 1991). In contrast, none of the differences between *R* homolog sequences are associated with long repeat sequence tracts, although several insertion/deletion mutations 3–69 bp in length are present in the variable regions.

The results of the molecular evolutionary analyses of both the grass *R* and the mammalian *SRY* sex-determining genes reveal that some regulatory loci may evolve by rapid diversification of much of their structure, while functionally required core domains remain conserved. Rapid evolution of large sequence tracts appear to be a recurring theme in regulatory gene evolution. Indeed, this mode of molecular evolution may increase the possibility of recruitment of these genes for new regulatory functions. Members of other eukaryotic regulatory gene families, such as homeodomain-containing proteins (SCOTT *et al.* 1989) or the plant MADS-domain developmental regulators (unpublished observations), show little sequence similarity to each other outside of DNA-binding and/or dimerization motifs. The finding that rapid sequence divergence occurs outside the bHLH domain in the plant *R* homologs suggests that accelerated molecular evolution within regulatory gene families may be one route by which novel developmental functions are established.

The authors wish to thank E. FRIAR, L. ARTHUR and J. BENNETZEN for providing DNA samples, R. D. DAMIANI and BETH ANDERSON for communicating unpublished data, and J. F. MCDONALD, R. MEAGHER and E. KELLOGG for valuable discussions. The authors would also like to thank KEN WOLFE and JOHN DOEBLEY for valuable suggestions, and W-H. LI for providing a copy of the program LI93. This work was funded in part by grants from the U.S. Department of Energy (to S.R.W.) and the Alfred P. Sloan Foundation (to M.D.P.).

LITERATURE CITED

AHN, S., and S. D. TANKSLEY, 1993 Comparative linkage maps of rice and maize genes. Proc. Natl. Acad. Sci. USA **90:** 7980–7984.

CREPET, W. L., and G. D. FELDMAN, 1991 The earliest remains of grasses in the fossil record. Am. J. Bot. **78:** 1010–1014.

DAVIS, R., H. WEINTRAUB and A. LASSER, 1987 Expression of a single transfected cDNA converts fibroblasts into myoblasts. Cell **51:** 1061–1067.

DELLAPORTA, S., J. WOOD and J. HICKS, 1983 A miniprep protocol for isolating plant DNA. Plant. Mol. Biol. Rep. **1:** 19–21.

DEPINHO, R., K. HATTON, A. TESFAYE, G. YANCOPOULOS and F. ALT, 1987 The human *myc* gene family: Structure and activity of *L-myc* and an *L-myc* pseudogene. Genes & Dev. **1:** 1311–1326.

DICKINSON, W., 1991 The evolution of regulatory genes and patterns in *Drosophila*. Evol. Biol **25:** 127–174.

DOEBLEY, J., 1993 Genetics, development and plant evolution. Curr. Opin. Genet. Dev. **3:** 865–872.

DOEBLEY, J., M. DURBIN, E. GOLENBERG, M. CLEGG and D. P. MA, 1990 Evolutionary analysis of *rbcL* nucleotide sequence among the grasses. Evolution **44:** 1097–1108.

ELLENBERGER, T., D. FASS, M. ARNAUD and S. C. HARRISON, 1994 Crystal structure of transcription factor E47-E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. **8:** 970–980.

EPPERSON, B. and M. T. CLEGG, 1987 Frequency-dependent variation for outcrossing rate among flower color morphs of *Ipomoea purpurea*. Evolution **41:** 1302–1311.

FERRE-D'AMARE, A., G. C. PRENDERGAST, E. B. ZIFF and S. K. BURLEY, 1993 Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature **363:** 38–44.

GOFF, S., K. C. CONE and V. L. CHANDLER, 1992 Functional analysis of the transcriptional activator encoded by the maize *B* gene–evidence for a direct functional interaction between two classes of regulatory proteins. Genes Dev. **6:** 864–875.

GOODRICH, J., R. CARPENTER and E. COEN, 1992 A common gene regulates pigmentation patterns in diverse plant species. Cell **68:** 955–964.

GOULD, S. J., 1977 *Ontogeny and Phylogeny*. Harvard University Press, Cambridge.

HELENTJARIS, T., D. WEBER and S. WRIGHT, 1988 Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. Genetics **118:** 353–363.

HILL, R. E., and N. HASTIE, 1987 Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature **326:** 96–99.

HUANG, N., G. L. STEBBINS and R. RODRIGUEZ, 1992 Classification and evolution of the alpha-amylase genes in plants. Proc. Natl. Acad. Sci. USA **89:** 7526–7530.

HUGHES, A., 1991 Circumsporozite proteins of malaria parasites: evidence for positive selection on immunogenic regions. Genetics **127:** 345–353.

HUGHES, A., and M. NEI, 1988 Patterns of nucleotide substitutions at MHC class I loci reveal overdominant selection. Nature **335:** 167–170.

LI, W. H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. J. Mol. Evol. **36:** 96–99.

LI, W. H., C. I. WU and C. L. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

LUDWIG, S., and S. R. WESSLER, 1990 The maize *R* gene family: tissue-specific helix-loop-helix proteins. Cell **62:** 849–851.

LUDWIG, S., L. HABERA, S. DELLAPORTA and S. R. WESSLER, 1989 *Lc*, a member of the maize *R* family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the *myc*-homology region. Proc. Natl. Acad. Sci. USA **86:** 7092–7096.

MARTIN, W., A. GIERL and H. SAEDLER 1989 Molecular evidence for Pre-Cretaceous angiosperm origins. Nature **339:** 46–48.

RADICELLA, P., D. TURKS and V. L. CHANDLER, 1991 Cloning and nucleotide sequence of a cDNA encoding *B-Peru*, a regulatory protein of the anthocyanin pathway in maize. Plant Mol. Biol. **17:** 127–130.

ROBBINS, T. J. CHEN, M. NORELL and S. L. DELLAPORTA 1989 Molecular and genetic analysis of the R and B loci in maize, pp. 105–114 in *The Genetics of Flavonoid Biosynthesis-Proceedings of a Post-Congress Meeting of the XVI International Congress of Genetics*, edited by D. E. STYLES, G. GAVAZZI and M. RACCHI. Edizioni Unicopli, Milano, Italy.

SCOTT, M. P., J. TANKUM and G. HARTZELL, 1989 The structure and function of the homeodomain. Biochim. Biophys. Acta **989:** 25–48.

STAPLETON, A., 1992 Ultraviolet radiation and plants: burning questions. Plant Cell **4:** 1353–1358.

SWOFFORD, D., 1993 *Phylogenetic Analysis Using Parsimony*. Illinois Natural History Survey, Champaign.

TANAKA, T., and M. NEI, 1989 Positive Darwinian selection observed at the variable region genes of immunoglobulins. Mol. Biol. Evol. **6:** 447–459.

TUCKER, P. K., and B. LUNDRIGAN, 1993 Rapid evolution of the sex-determinig loci in Old World mice and rats. Nature **364:** 715–717.

VARAGONA, M. J., M. D. PURUGGANAN and S. R. WESSLER, 1992 Alternative splicing induced by insertion of retrotransposons in the maize *waxy* gene. Plant Cell **4:** 811–820.

WHITFIELD, L., R. LOVELL-BADGE and P. GOODFELLOW, 1993 Rapid sequence evolution of the sex-determining gene *SRY*. Nature **364:** 713–715.

WOLFE, K., M. GOUY, Y. YANG, P. SHARP and W. H. LI, 1989 Date of the monocot dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA **86:** 5201–5205.

Communicating editor: W-H. LI